To: CN=Stephanie Santell/OU=DC/O=USEPA/C=US@EPA[]
Cc: []
From: CN=Tina Laidlaw/OU=MO/OU=R8/O=USEPA/C=US
Sent: Tue 10/9/2012 5:07:12 PM
Subject: Fw: Yellowstone River NSTEPS Comments
DEQ_NSTEPS_Responses_10-5-2012.pdf

Stephanie,

Haven't read these comments but wanted to forward them along.

Tina

Tina Laidlaw
USEPA Montana Office
10 West 15th Street, Suite 3200
Helena, MT 59626
406-457-5016

----- Forwarded by Tina Laidlaw/MO/R8/USEPA/US on 10/09/2012 10:36 AM -----

From: "Flynn, Kyle" <KFlynn@mt.gov>
To: Tina Laidlaw/MO/R8/USEPA/US@EPA
Cc: "Suplee, Mike" <msuplee@mt.gov>, "Pipp, Michael" <mpipp@mt.gov>, "Urban, Eric"
<EUrban@mt.gov>
Date: 10/05/2012 03:57 PM
Subject: Yellowstone River NSTEPS Comments

Hi Tina –

Attached are DEQ's responses to the NSTEPS peer review for the Yellowstone River work (in electronic
form). Please let me know if you would like any additional information or clarification, or if you would like
me to send a hard-copy via mail.

Best regards,

Kyle Flynn, P.H.
Lead Hydrologist
Montana Department of Environmental Quality
Water Quality Modeling Program
Water Quality Planning Bureau
1520 East 6th Ave.
Helena, MT 59620
Tel:(406) 444-5974
Fax:(406) 444-6836

October 5, 2012

Tina Laidlaw
USEPA Montana Office
10 West 15th Street, Suite 3200
Helena, MT 59626

Hi Tina:

Enclosed is a memo containing Montana Department of Environmental Quality's (DEQ's) responses to the Nutrient Scientific Technical Exchange Partnership & Support (NSTEPS) peer review for the Yellowstone River nutrient criteria model. We have done our best to address each comment (where appropriate) and will be revising the draft report accordingly. In an effort to make the subsequent pages easy to follow, we have shown the reviewer's comment in italics and our response in plain text. Please let us know if you need clarification, or additional information about any of the content.

Finally, we apologize about the lengthy turnover time in our response. This was largely a function of my academic commitments over the last year. In any regard, we look forward to discussing items as needed.

Best wishes,

Kyle Flynn, P.H.
Lead Hydrologist
Montana Department of Environmental Quality
Water Quality Modeling Program
1520 East 6th Ave.
Helena, MT 59620
Tel:(406) 444-5974
Fax:(406) 444-6836

Enforcement Division   ●   Permitting & Compliance Division   ●   Planning, Prevention & Assistance Division   ●   Remediation Division

0015227

SECTION 1.0 - Responses to Reviewer 1

"General Comments
This is a well written report on "Using a computer model to derive numeric nutrient criteria."
There are relatively few errors in the draft, which made reviewing clear. The use of multiple sources of information, including a computer model, is a very good idea for establishing nutrient criteria. The many concepts developed and employed in this effort are innovative, well founded, and sound. However, I disagree with the conclusions that model conditions warrant more credibility than other sources of information and that model results should be used to set nutrient criteria for the Yellowstone River.

In summary, my short responses to the questions are:
1. *The data used to run, calibrate, and validate the model were appropriate, but not sufficient.*
2. *Model calibration and validation were not good, because the fit of data to model runs was poor for a key endpoint variable, benthic algal biomass, and many results were biased.*
3. *The uncertainty of model predictions was problematic because: the model was not validated well for a key endpoint variable; the model was used to extrapolate to nutrient conditions outside the range for which it was calibrated and validated; and the model did not simulate extreme values well.*
4. *pH and algal biomass response endpoints should be used to establish nutrient criteria. The most sensitive response to a stressor (i.e. nutrients in this case) should be used to establish stressor criteria, even if different response endpoints are most sensitive in different types of habitats (in this case shallow and deep river habitats).*
5. *The appropriate methods were used to gather information about the development of nutrient criteria, but the results of the computer model were overstated and overweighted in a premature decision on nutrient criteria."*

1. Please evaluate the sufficiency and appropriateness of the data used to run the model.

"The data used to develop the model was appropriate, but not sufficient.

The computer model was designed to measure important response variables, such as benthic algal biomass, pH, and DO. These parameters respond either directly or indirectly to variation in nutrient concentrations and are used in either narrative or numeric water quality criteria in many states. These variables are highly appropriate from the perspective that we want to protect uses of waters. We know enough about nutrients to know the effects of nutrients instream and downstream. With proper research and synthesis of results, we should be able to set nutrient criteria above minimally disturbed conditions without threatening designated uses, such as drinking water, recreational uses and aesthetics, and support of biodiversity. Although we may not be protecting aquatic biodiversity of taxa that are highly sensitive to moderately increased nutrient concentrations in a habitat with nutrients above minimally disturbed condition, presumably those taxa are being protected in other habitats in which minimally disturbed condition is being protected (invoking tiered aquatic life uses). With the knowledge that biodiversity of some nutrient sensitive taxa will not be protected at nutrient concentrations that generate algal biomasses greater than 150 mg chl a m^{-2} and pH and DO standard violations, benthic algal biomass, DO, and pH can be appropriate endpoints for managing nutrients."

We disagree with the first portion of this comment (i.e., *"The data used to develop the model was appropriate, but not sufficient"*) and suggest that the DEQ effort meets/exceeds most steady-state modeling applications (see Mills et al. 1986; Barnwell et al. 2004; and reviewer 2's comments), including prior modeling studies in the literature (Paschal and Mueller, 1991; Park and Lee, 2002; Kannel et al. 2006; Turner et al. 2009). If anything, we feel it should be described as comprehensive.

0015229

"The right variables were modeled, measured, and calibrated in the field, but the sample size was low. Many of the key environmental variables were measured in the field, but they were measured at less than 10 locations. This limits the power of the comparison, much as a low sample size limits the statistical power in hypothesis testing. Was the fit or the lack of fit of the model to data due to chance or was it true?"

Sample size is just one of several factors that should be considered in modeling. According to Mills et al. (1986), other factors include site accessibility, historical locations, critical points of maximum or minimum concentration, and locations where water quality standards are expected to be violated. Because there are no hard and fast rules for sample size, an appropriate n is left up to the professional judgment of the modeler. Mills et al. (1986) suggest the sample size should be sufficient to describe the longitudinal profile of the river. So in the case of the Yellowstone, this was done. For example, we accommodated variability such as incoming tributaries, wastewater treatment plant discharges, critical downstream points of concentration, and spatial differences in temperature brought about by climatic gradients and hydrogeomorphology. So for the reviewer to suggest that random chance explained the structural differences in the data (e.g., larger diel oxygen swings in enriched areas, changes in algal biomass, increasing suspended solids, etc.), is simply not plausible. In this regard, we find the reviewer's comment speculative and without basis.

"The study should have been designed to have the calibration and validation datasets at the same time of year, perhaps sampling during summers of 2007 and 2008. The differences in temperature and light (day length and sun angle) between August and September could be substantial given they are within range that macroalgae like Cladophora are especially sensitive. August and September also have very different algal accumulation histories and processes regulating algal ecology probably differ as a result. Interannual variation in physical and chemical conditions in the Yellowstone River are relatively predictable, because of discharge regulation by snowpack melting, compared to rivers in parts of the country where unpredictable rain events have great effects on discharge and resulting physical and chemical conditions (e.g. light and nutrient concentrations)."

Similarity of environmental conditions (e.g., light, temperature, etc.) is not a necessity when considering mechanistic studies. Process-based models explicitly account for water temperature variation, solar radiation/time of year, biological rates, etc. thereby accommodating the differences pointed out by the reviewer. In fact, Chapra (2003) actually suggest that process-based models be calibrated and validated to substantially different conditions, such as flow, loadings, or climate. For example, a Level 2 model confirmation (i.e., the best) would require the model to be applied to cases with significantly different loadings and meteorology. While we did not meet this stringency, we did achieve a Level 1 confirmation which essentially means the model was applied to different meteorology and flows. That said, the accumulation history/autocorrelation of algae between August and September is a valid concern. We are currently investigating whether this is an important consideration or not.

"Another concern was having sufficient scientific foundation for model coefficients. Admittedly, some knowledge is better than none, but assuming that coefficients developed in lakes or other parts of the country and for different kinds of algae in one condition or another would apply to this location seems premature. Many of the parameters were developed in the 1970s or earlier, not that old is necessarily bad, but it is an indication that few new components were available or were found in the literature for use in the computer model. More field and laboratory research is needed to quantify the parameters being used in processed based models."

We did <u>not</u> directly apply coefficients from lakes or other parts of the country as suggested by the reviewer. Rather we made an initial assumption about such values (and associated ranges from the literature) and then calibrated those values to site-specific measurements (e.g., biomass, chemistry, water quality data, etc.). Such practice is common in water quality modeling and eliminates the need for direct parameter transfer as suggested by the reviewer. So the real issue seems to be kinetic parameterization of the model. We can only point to the fact that we used a combination of field/laboratory studies (e.g., light-dark bottle experiments, delta-method, SOD measurement, etc.) and field-calibrated state-variables (e.g., DO, pH, algal biomass, etc.) to provide the best (admittedly not perfect) model representation. Allowable ranges of coefficients were bounded by the literature and included quantification of both parameter sensitivity and uncertainty through first-order error and Monte-Carlo analysis techniques. While we agree that more data is always nice (note: we would love to do more field and laboratory research), at this time enough is known about site-specific biogeochemical processes (e.g., algal assimilation, hydrogeometric properties, chemical kinetics, etc) to provide reasonable assessment of the river's eutrophication response for regulatory purposes.

2. Please evaluate the model calibration and validation

"Model calibration and validation were not good, because the fit of data to model runs was poor for a key endpoint variable, benthic algal biomass, and many results were biased."

It is unclear to us what "not good" is, but root mean squared error (RMSE) of our simulation was 21.8 and 35.0 mg Chl$a$ m^{-2} during August and September 2007 (n=77, excluding filamentous sites and a site with nitrogen fixers). Using a worse-case combination which includes filamentous algae and nitrogen fixers, RMSE was 55.5 mg Chla/m^2 (n=90), which approximates a seasonal average (i.e., average of August and September). While such errors are apparently large (according to the reviewer) they are no worse than routinely reported for empirical studies in the literature. For example, we compiled regression statistics via digitization of figures for about a half dozen of the more commonly cited nutrient-algal biomass papers and found that benthic algal biomass predictions, whether empirical or mechanistic, are quite similar (**Table-1**). In fact, the mechanistic model performed slightly better in nearly all instances than the studies considered. Plus it had the added benefit that other water quality state-variables such as DO, pH, etc. could also be simulated which cannot be done with a simple biomass model.

In consideration of **Table-1** though, it is important to keep in mind that the relative magnitude of RMSE is influenced by the range of biomasses evaluated, i.e., larger biomasses have the potential for greater prediction error than smaller biomasses and thus artificially weight the computed RMSE statistic. Thus some caution is needed in interpretation of results. Likewise, we suggest a more thorough review of both mechanistic and empirical models be completed before a definitive conclusion can be made about the predictive ability of each model type.

Finally, as pointed out by the reviewer, our model does contain bias. We have described it in Section 10.4.3.2 as under-prediction of high biomass and over-prediction of lower biomass (especially for filamentous algae). The prediction problems at the upper end reflect the inability of the model to simulate filamentous growth whereas those at the lower end are strictly applicable to diatom species. We clearly would like to remedy this deficiency, however, given the amount of filamentous algae in the lower Yellowstone River, further time and resource spent on model development is not

0015231

warranted. We will address the filamentous concerns in the future, when both algal communities are present and necessitate the development of a model with better prediction capability.

Table 1. Comparative error analysis of commonly cited literature studies.

Study	Location	RMSE (mgChla/m²)	n
Lohman et al. (1992)	12 streams and 22 sites in northern Ozarks, Missouri (annual mean of TN)	27.4	44
This study	90 algal sites Yellowstone River, Montana (instantaneous measurements during growing season)	29.6[1] 55.5[2]	77 90
Dodds et al. (1997)	205 streams or sites worldwide (seasonal mean of TN)	49.5	146
Suplee et al. (2012)	8 sites Clark Fork River, Montana (seasonal mean TP)	73	84
Chételat et al. (1999)	13 rivers in southern Ontario/western Quebec (TP)	85.4	33
Biggs (2000)	25 runoff fed rivers in New Zealand (SIN)	326.5	30
Welch et al. (1992)	26 sites in 7 New Zealand streams; mechanistic model	723	26

[1] Excluding sites where filamentous biomass or nitrogen fixers were present.
[2] All sites.

"Not much change was needed in many model parameters to calibrate the model, but many parameters for benthic algal growth were substantially different between the initial estimate and calibrated value (Tables 9-5, 9-6, and 9-7). Almost no discussion followed on the magnitude of these changes and if they were reasonable."

Initial parameter estimates are based on previous recommendations or initial data evaluations which must be adjusted on a per-system basis through model calibration (as described previously). Thus the magnitude of change from the initial parameter estimate is not a factor of whether a calibration is suitable or not (the fit between the observed and simulated data is!). In retrospect, we could have probably done a better job describing this in the text though. We did provide details on where estimates originated from in Section 8 (e.g., C:N:P ratios, subsistence quotas, nutrient uptake estimates, etc.) and we will be sure to add this reference to Section 9. Finally, we will add text describing the fact that values must be calibrated (i.e., an initial estimate is just that, and deviation from that does is not a significant concern provided the calibrated value is within the range of the literature).

"At least one set of the changes in parameters was relatively easy to evaluate and determine if they were reasonable. The mass ratio of N:P in algal cells is assumed to be 7:1, and in the Yellowstone River was often lower because of the relatively low supply of N versus P in the river. The initial mg N and P per mg algae (subsistence quotas for N and P) for benthic algae were assumed to be 0.7 and 0.1, respectively (Table 9-6).

- *The real issue is the relatively large change in one value during calibration and the unrealistic ratio for parameter values resulting from that calibration. The resulting calibration values of parameters for subsistence quotas for N and P were 3.20 mg N and 0.13 mg P, respectively. Even though each of these parameters independently fit within the range of possible values reported in the literature (remembering that one outlier in the literature has great effects on this range), the ratio seems very high for conditions within the Yellowstone River. The resulting mass ratio of subsistence levels of N and P was 3.20:0.13, which is more than 3 times the expected 7:1 ratio and 6 times the 4:1 ratios observed in low N habitats like the Yellowstone."*

0015232

It is commonly misconceived that subsistence quotas scale at Redfield ratio (7.2:1 by mass). However, Shuter (1978) provides a compilation of minimum cell quota data for N and P vs. biovolume (for phytoplankton) that seem to disprove this. From data on more than 25 algal species it is shown that N to P ratios deviate substantially from Redfield near the minimum cell quota. Recent work by Klausmeier et al. (2004) supports this assertion. They suggest resource acquisition machinery (i.e., nutrient-uptake proteins and chloroplasts) are P-poor, making the N:P ratio higher (ca. 20-30:1 by mass) nearer to the cell quota. Conversely, under nutrient replete conditions (more like Redfield) P-rich ribosome assembly machinery for exponential growth is more prevalent leading to lower N:P ratios. All of these findings are consistent with the classic work by Goldman et al. (1979) where it is shown that algal cellular N:P ratios are strongly influenced by the alga's growth rate. At very low growth rates (i.e., those approaching the minimum cell quota) cellular N:P ratios increase greatly to 45:1 (by mass). Hence we feel the ratio we have in the model is justified.

- *"Although internal N and P half-saturation constants are substantially different types of parameters than subsistence quotas, both are involved with algal growth, both were changed substantially during calibration, and ratios for both were unusually high."*

Very little data exists on internal N and P half saturation constants so we assume that this comment is pertaining to the external values. As mentioned previously, deviation from the initial estimates is not a problem (referring back to our previous response to this same question). However, we do agree the values required for calibration seem high in comparison to other work (e.g., Bothwell; 1985, Borchardt, 1996; Rier and Stevenson, 2006). That said Bothwell (1989) shows that low saturating levels are probably only valid during the cellular growth, at a time when nutrient supply is high and is not impeded by diffusion through the algal mat. Thus when algal biomasses are higher (or detrital accumulation is significant), it is possible that nutrient gradient/diffusion limits nutrient supply which may explain why higher values are needed to calibrate the model to a natural river. It is important to also realize that the Droop (1974) internal stores model is being used and thus to frame the overall response as a Michaelis-Menton or Monod saturation model, output biomass and soluble nutrient levels must be considered. By doing this we found that peak biomass saturated at around 152 µg/L soluble inorganic nitrogen (SIN) and 48 µg/L SRP (when not limited by other factors). Values such as these are not that different than suggested by the literature thereby providing additional confidence in the model's predictions.

Note: If the comment was specifically about internal half-saturation constants (the capacity for nutrient uptake based internal cellular stores), we acknowledge these values are poorly understood. Our best understanding is that they can be scaled in accordance with subsistence quotas at a ratio of around 1.0 for N and 0.5 for P (Di Toro, 1980; Droop, 1974; Rhee, 1973; Rhee, 1978). Given the uncertainty in their value, they were calibrated.

- *"The same kinds of problems were noted for the phytoplankton (Table 9-7)".*

Again, initial phytoplankton coefficients are estimates only, and must be calibrated. We will add a discussion regarding deviation from the initial estimates and what this means.

- *"A confusing issue initial parameter values (e.g. 0.7 mg N or 0.1 mg P per mg algae) indicate 70 and 10% of the algae were composed of N and P. Most of algal mass is carbon, not N or P. Presumably the units or my understanding of what these parameters mean were wrong."*

The reviewer is correct that the units could be easily confused. The values referenced are the initial estimates of minimum cell quota, or minimum level of nutrient deficiency normalized to Chl*a* [i.e., before our review of the Shuter (1978) or Klausmeier et al. (2004)]. As suggested by the reviewer, the actual makeup of algal cells is much different at a stoichiometric ratio of 40 mgC to 7.2 mgN to 1 mgP (i.e. Redfield).

"Fit of the model, similarity between predicted and observed conditions, was better for physical than chemical parameters, and better for chemical than biological parameters. QAPP criteria were not met for 1 out of 5 of the parameters assessed (Table 10-1). The variable with poor fit based on RMSE and RE was benthic algal biomass, either by using the Q2K or AT2K model. Since benthic algal biomass was a key response endpoint, and an endpoint for which nutrient criteria were eventually going to be made, it was important that the model predict benthic algal biomass well."

This is correct, the poorest part of the simulation was the biological component. However, the algal simulation error was quantified and was no worse than if we were to use other methods [referring to the previous discussion about Lohman et al. (1992), Dodds et al. (1997), Chetelat et al. (1999), Biggs (2000), etc.]. So if past efforts were acceptable (some of which were used in criteria determination), why would this effort be any different?

"As suggested on page 10-21, I agree that the AT2K model "allows us the ability to gain better information about spatial relationship of biomasses across a river transect," but I don't agree that AT2K model predictions were sufficiently accurate for the purposes intended for the modeling effort. High benthic algal biomasses were consistently under-predicted."

As indicated previously, the model's accuracy is comparable with past studies which means it should be suitable for its intended purpose (i.e., nutrient regulation on large river during the growing season where a vast majority of algal growth is closely attached to the bottom). That said, long isolated streamers of filamentous algae such as *Cladophora* present a problem. Computed biomass is greatly underestimated in these instances and we attribute this to the fact that the model simulates benthic growth in one-dimension vertically (i.e. thickening of an algal mat). In contrast, long *Cladophora* streamers grow up into the water column in 3-dimensionally which results in considerably higher biomasses for a given nutrient level and spatial area. Fortunately about 97% of all algal samples were diatom-like, so we do not see the underprediction of these isolated instances an issue (note: species shifts from diatoms to filamentous are a valid concern and we will evaluate this consideration if the river moves closer to the established criteria).

"During review of figures, I became concerned that deviations between observed conditions and conditions predicted by the model are more serious if they are biased than if they are randomly distributed above and below model predictions. This bias would not be captured in the RMSE and RE statistics for goodness of fit. For example, even though the RE is only 7.3% for TN calibration and 1.38% for validation (Figure 10-7, the model overestimates TN concentrations). The bias in predictions (residual error) is common in many of the nutrient and biological parameters. In most cases, bias was either high or low along the river, but in some cases it systematically switched from high to low, which you could imagine was the case for the August 2000 phytoplankton validation

(Figure 11-9). Systematic bias along the river is a concern because habitat conditions change systematically along the river."

We agree with the reviewer that model bias is undesirable. However, the level of bias suggested (ca. 10%), is hardly of concern (see Moriasi et al. 2007). Errors of this magnitude are considered "good" in the modeling literature. More importantly we feel the reviewer is mistaken in characterization of error calculation. RE is in fact a direct measure of bias, e.g., it sums the residual errors (predicted-observed) and divides those by the observations. So for the figure of concern (i.e., Figure 10-7), approximately 50 µg/L of bias occurs. While such an error is not conservative (i.e., does not side with the resource) this is not a great concern given the overall magnitude of nutrient levels in the river. Also, from review of the summary statistics in Table 10-1, it should be noted that several state-variables have larger bias. These are detailed in subsequent comments. Finally, with systematic bias, we would suggest this has more to do with data variability than systematic model error. While systematic habitat changes do occur in the river (e.g., shallowing near Miles City, increased turbidity below the Powder River, water temperature changes, etc.), we have characterized these features well and do not see how systematic artifacts could occur so rapidly in the longitudinal profile (referring to the reviewer's contention about the August 2000 phytoplankton data).

The model did not capture extreme conditions well, especially for benthic algae. If there was little variation, the model tended to fit much better than if a parameter varied greatly over the range of nutrient and habitat conditions in the river. For example, diurnal variation in dissolved oxygen and discharge were simulated well by the model, but pH and benthic algal biomass which varied much more than DO and discharge were not simulated well by the model.

The model may not have been able to simulate the high algal biomasses that accumulate in the river. For example in Figure 10-15, the model never predicted algal biomass to be greater than about 70 mg chl a m^{-2}. However, several observations of higher chlorophyll were observed. In addition, most of the observed levels of chlorophyll a were less than 50 mg chl a m^{-2} and fell within a confidence envelop that probably had a width of 40 mg chl a m^{-2}. So it would have been difficult for the model to be wrong when benthic algal biomass was less than 50 mg chl a m^{-2}. When benthic algal biomass was predicted or observed to be greater than 50 mg chl a m^{-2}, only 1 of the 10 prediction/observation points were within the RMSE confidence envelop."

In regard to the benthic algae simulation (and the inability to simulate high biomasses), the reviewer is correct that the cumulative frequency plot in Figure 10-15 shows under-prediction of higher biomasses which is a concern to us as well. We have been forthcoming about this in our discussion, and did additional analysis to make certain that the model would generate anticipated biomass levels under eutrophied conditions. This is described in Section 8 and Figure 8-5 and we show that maximum expected biomasses under nutrient and light replete conditions (with assumed losses of 50% from respiration and scour/grazing) would be around 300-400 mgCh*la* m^2, similar to that suggested by Stevenson, et al. (1996) for diatom communities. So while the model did consistently underestimate some field measurements (mostly filamentous algae), it will achieve maximum expected diatom community biomass under nutrient enriched conditions. Finally the reviewer is technically correct that the RMSE envelope covers nearly the entire simulated range (i.e., in their comment *"it would be difficult for the model to be wrong"*). However, this comment is somewhat misleading as nearly all of the data falls along the 1:1 line (in a structured fashion) and is certainly not random as inferred by the reviewer.

0015235

"Another issue with this model fit analysis is also the skewness of the distribution of observed and predicted values, with most points within 1/6th of the range of potential values (<50 mg chl a m^{-2} with a range of 0-300 mg chl a m^{-2}). Basically, it seems the model was not tested in the range of conditions in which it is intended to be applied."

We have no control over the skewness of the data as it is simply a function of field conditions and data collection methodology. The reality is that given the nutrient and light limitation of the river biomasses are low (<70 mgChl$a$/m^2), with exception of a few anomalous filamentous algal point measurements. With this understanding, it is surprising to us that the reviewer suggests we failed to test the model over the range of appropriate conditions. The immediate question that comes to mind is: (1) would we need a model if such conditions were already occurring and (2) could river-wide conditions for everything else (DO, pH, nutrients, etc.) be reasonably determined using any other approach (e.g., such as experimental troughs)? The obvious answer to both is no. Hence the primary purpose of the model is to help understand the response to a given set of enriched conditions while at the same time maintaining the fundamental/theoretical constructs of the eutrophication process. Finally, the reviewer is incorrect when implying that empirical restrictions be placed on process-based models. It is well-known that mechanistic models are a useful for predicting conditions outside of the environmental conditions they were developed (EPA, 2001; Canham et al., 2003).

3. Please comment on the uncertainty in the model predictions

"The uncertainty of model predictions was problematic because: the model was not validated well for a key endpoint variable; the model was used to make predictions for nutrient conditions outside the range for which the model was calibrated and validated; and the model did not simulate extreme values well. In particular, the inability of the computer model to simulate extreme values in benthic algal biomass was a concern."

We tend to disagree with this blanket statement and have described why in previous responses. To reiterate: (1) we <u>did</u> show that the algal simulation was no worse (in fact better) than many of the literature suggested approaches, (2) contrary to what the reviewer has indicated, it is OK to apply a mechanistic model beyond conditions which it was calibrated/validated (provided assumptions used in development of the model are valid), and (3) simulating extreme values (i.e., isolated cases where filamentous algae occur) is not an important consideration in this study.

"The poor prediction of algal biomass and inability to really evaluate model prediction of pH and other important response variables was discussed above."

The reviewer has not anywhere demonstrated a deficiency to evaluate pH or other important response variables (such as DO, nutrients, etc.). The fact is, short of benthic algae (which seems to be the reviewer's main focus), nearly all simulated state-variables achieved QAPP project requirements (and even algae did in one instance).

"A basic tenet of modeling, either statistical or highly calibrated computer models, is limiting extrapolation of results outside the range of conditions in which the model was developed. This model was employed outside the range of conditions for which it was calibrated. Since the computer model performed much worse when applied to September than August conditions, due to likely seasonal effects, wouldn't we also expect the same issues with performance outside the range of nutrient concentrations in which the model was calibrated?"

The reviewer's statement regarding extrapolation of modeling results conflicts with EPA guidance. In fact, EPA (2001) clearly articulates in Chapter 9, Use of Models in Nutrient Criteria Development that, *"Considerably more space is devoted to mathematical models, because they are capable of addressing many more details of underlying processes when properly calibrated and validated. They also tend to be more useful forecasting (extrapolation) tools than simpler models* (referring to empirical models)*, because they tend to include a greater representation of the physics, chemistry, and biology of the physical system being modeled (NRC 2000)"*. We therefore do not understand the reviewer's concern, especially since process-based models have a long and successful history in waste-load allocations and effluent loading studies (Thomann, 1998; Chapra, 2011).

With respect to the seasonal issue (September vs. August), there is no reason to make the linkage suggested by the reviewer. We in fact provided a very satisfactory explanation for the deficiency between August and September 2007 and also completed a second validation for August 2000 which confirms the model performs well during peak growth conditions (i.e. August). Additionally, the calibration and confirmation were collectively completed over a range of different soluble nutrient conditions including nitrogen levels ranging from 5-105 µg/L and phosphorus concentrations from 3-17 µg/L (across the longitudinal profile). As such, soluble nutrients spanned almost the entire range evaluated for criteria determination, with the caveat that nutrient supply was elevated over only a small spatial extent usually in the vicinity of the wastewater treatment plants. Thus to question the model performance over a period which in essence has already been validated for varying nutrient conditions (i.e., August) is unjustified.

"Process based models (i.e. computer models) are theoretically better than statistical models for predicting outside the range of original conditions in which they were calibrated. However, the extent and magnitude of calibration from an initial values used in model is a key issue for using process based models to predict outside the range of calibration. Prediction outside the range of conditions for which either the statistical or process based model was calibrated requires that we know enough about the system and the behavior of the system in the two ranges of conditions (e.g. August versus September, or low and high nutrient concentrations) that we are confident that the models accurately describe behavior of the system. The less that you have to calibrate a model to new conditions to get a good fit, the more confident you can be that the model will perform well in a new set of conditions. The more fundamental the processes are that are simulated in the model and the fewer number of assumptions made for use of the model, the more certain you can be that the model will predict responses well in a set of conditions for which it was not calibrated.

Since there is little evidence that the model did perform well, either calibrating for key endpoints or predicting responses during validation, we should have concerns about accuracy of predictions by the model for ecological responses in higher nutrient concentrations for which the model was tested. In addition, many key parameters in the model were changed greatly during calibration from what were initially thought to be appropriate. So based on model performance, we cannot be certain that it will perform well outside the range of conditions in which it was calibrated, or even within that calibration range for some key parameters."

We agree that process-based models are better than statistical models for predicting conditions outside the range which they were developed (i.e., that is their primary utility), but disagree that *"there is little evidence that the model did perform well"*. If fact, we have clearly articulated the model's predication capability throughout the draft report as well as in many of our responses. One further clarification is necessary though. The reviewer describes August and September as *"low and*

high nutrient concentrations". However, this is not the case. Rather nutrient supply was the same both periods (i.e. loadings were similar), but uptake during each period was significantly different. Finally, with respect to the certainty of model predictions, the entire premise of the model is to represent fundamental biogeochemical processes. These were shown to be adequate for August low-flow conditions (based on two different years of data, i.e., 2001 and 2007) and over a large longitudinal extent. Thus it is reasonable to conclude that the model is suitable for making regulatory predictions over this time-frame, especially since as noted previously, nutrient supply was sufficiently variable in both years.

"Many assumptions needed for the model also seemed to reduce credibility of its results. Some assumptions were probably met as well in the Yellowstone River as anywhere. For example, the assumption about the model simulating a steady state equilibrium is certainly more appropriate for rivers like the Yellowstone with snow-melt dominated and relatively predictable hydroperiods versus many other rivers where storm events have dramatic and unpredictable effects on hydroperiod."

Violation of model assumptions by the ecosystem may also explain why the model simulated the ecosystem poorly. Of course assumptions are necessary, but some violations of assumptions or combinations of violations may accumulate explain the unsatisfactory behavior in the model. Here are a few examples:

- *The assumption that velocity and channel substratum are "sufficiently well mixed vertically and laterally" (pg 5-8, lines 3-4) may explain why the high algal biomasses were not simulated. If average, versus optimal velocity and substratum were used that would underestimate the high algal accrual possible in optimal velocity and substratum conditions."*

We disagree that the model simulated the ecosystem poorly (for all of the reasons stated previously) but do agree that spatial variability of substratum and velocity may be an important consideration in algal growth. We are working on improving modeling techniques to better represent these physical processes in riverine settings. The assumption of vertical and lateral mixing referenced by the reviewer holds only for the water column (i.e., turbidity, nutrient concentrations, phytoplankton, etc) and we will revise the text to make this clearer.

- *"Why assume dynamic equilibrium between particle re-suspension (drift) and deposition (settling)(pg. 8-20, lines 24-25)?"*

We will rewrite this sentence to clarify. Dynamic equilibrium between particle resuspension and settling was based on conclusions of Whiting, et al. (2005) which was based on longitudinal sediment analysis of the Yellowstone River. For the model we applied our calculated Stokes settling velocity of 0.012 m d^{-1} for sediment and 0.086 m day^{-1} for phytoplankton (calibrated down to 0.05 m day^{-1}) reflecting a net loss in the mass balance for each term.

- *"Why assume the typical meterological year during a ten year period. For example, to understand the conditions under which problems would arise 1 in 10 years, aren't regional weather patterns a likely cause of those problems. Rather than running a typical meteorological year, shouldn't the 10-year extremes be boundary conditions for a run to understand the effects of less common conditions?"*

0015238

The use of a typical meteorological year stems from the desire to not alter the underlying frequency of occurrence. For example, if a 1 in 10 year low-flow condition were simulated with a 1 in 10 year climate (both of which have independent probabilities), the underlying design condition would be a 100 year event (probability of occurrence of 0.1 x 0.1 = 0.01). Such an infrequent event is <u>not</u> appropriate for nutrient regulatory management. As indicated by the reviewer, however, an equally viable approach would be to use a 1 in 10 year climate, with a 1-year flow condition although in this instance the latter reflects a much larger system volume (from the increase in flow) which would likely outweigh any extreme climatic effects.

Note: We have modified the design flow to a 14Q5 (1 in 5 year low-flow condition) to better align with EPA recommendations on allowable frequency of exceedance of standards (which were originally based on a biologically 4-day average flow once every 3 years, i.e., 4B3). The 4B3 is often used as a basis for U.S. EPA chronic aquatic life criteria.

"In addition to violation of the assumptions in the model, there may be issues with the analytical foundation of the model to accurately represent ecosystem processes; but I am not sufficiently familiar with the model to make that judgment. For example:

- *Were growth patterns and differing spatial resource limitation (density dependence) for macroalgae and microalgae or algal taxa included in the model?"*

It would have been helpful for the reviewer to familiarize themselves with the model prior to doing a critique of its analytical foundation, but in general we will try to answer each question straightforwardly. Relative to different growth patterns/state-variables for each algal taxa, Q2K models only a single algal species therefore any difference between macro and micro-algae species is only accommodated through parameter lumping. We recognize this as a model deficiency (especially if applied in an area where both macro and micro-algae were in competition), however, a majority of the river sampling sites (~97%) were dominated by a mixed assemblage of diatom species which at least reduces the concern of macro- and micro-algal dynamics. Thus it was not a concern in the modeling endeavor.

- *"Space limitation in the model, if I understand it correctly, is not the correct conceptualization of the process that regulates density dependent growth of benthic algae. Developing a more realistic characterization of the processes regulating benthic algal accumulation and density-dependent depletion of nutrients within mats would be very interesting and perhaps improve model predictions. Effects of mixing and diffusion vary greatly between different types of algae that grow in differing nutrient and temperature ranges, such as macroalge (Cladophora) and microalgae (diatoms)."*

While in one section of the report we use a logistic function/space limitation to illustrate biomass accumulation for the purpose of estimating zero-order growth rates (under optimal nutrient and light conditions), such a formation is not actually used in the Yellowstone River model. Instead the governing differential equation for the mass balance of algal biomass is based on Chapra et al. (2008) where biomass increases due to photosynthesis and is moderated by a number of loss terms including respiration, excretion, and death (inclusive of grazing and scour etc.). This would have been clear if the reviewer would have taken the time to review the Q2K model which can readily be found on the EPA website http://epa.gov/athens/wwqtsc/html/qual2k.html. The

model is based on the work of McIntire (1973), Horner et al. (1983), Uehlinger, (1996), and Rutherford et al. (2000), includes Droop (1974) nutrient limitation (i.e., the internal stores model), saturation light limitation (Baly, 1935; Smith, 1936; Steele, 1962; light), uptake dependent on internal and external nutrients (Rhee, 1973), and many other physiology-based processes. In this regard, the effects of nutrient diffusion into the algal mat are not explicitly considered, but are implicit in calibration of the external half-saturation constants for nutrient uptake.

- *"Was N-fixation included in the model and the potential for N transfer between epiphytic diatoms with cyanobacterial endosymbionts on Cladophora? Is it possible that Cladophora cells close to the substratum take up nutrients and transfer them to younger, actively growing cells in the ends of the filaments suspended in the water column. Only the cells at the tips of Cladophora filaments reproduce, so they are younger and have fewer epiphytes than cells at the base of filaments. Cladophora cells that are closer to the substratum, having more epiphytes, bacteria, and entrained detritus as well as slower currents, have greater potential for uptake of recycled nutrients in the epiphytic assemblages around them than younger cells in the water column. Cladophora does not have complete cross walls between cells, so fluid in cells can theoretically mix between cells, which would be facilitated by the movement and bending of filaments in currents. Thus, nutrient concentrations in the water column may be poor estimators of nutrient availability to Cladophora, as well as other benthic algae, because of nutrient entrainment and recycling in the mats."*

N-fixation is not included in the model and its importance (at one site) was identified only after finding discrepancies between simulated and observed data. Similarly, nutrient exchange from epiphytic diatoms with cyanobacterial endosymbionts to *Cladophora* is not represented. Both are far too detailed processes for a general purpose water quality model. Finally, while the *Cladophora* mat self-sustainment process described by the reviewer is interesting and may occur, the concept seems in conflict with the common observation in Montana and elsewhere that dense stands of long streamers of *Cladophora* most frequently colonize the riffle regions of streams and rivers; this was reported as long ago as 1906 (Fritsch, 1906). Increased turbulence and advection in riffles clearly creates preferred habitat, in part because it induces more nutrients from the water column to go deeper into the mat, allowing for continued photosynthesis (Dodds, 1991). If the mat nutrient-recycling process described previously is important to mat maintenance, there is still the obvious question of what stimulates *Cladophora* mats and long streamers to develop in the first place? The scientific literature is replete with works dating back to at least the 1950s indicating that *Cladophora* blooms are associated with elevated nitrogen and phosphorus concentrations in the water of rivers and streams (see Whitton, 1971 and Hynes, 1966 for starters). As such, we believe the scientific literature generally supports the idea that nutrient concentrations in flowing waters are correlated with the development of algal mats.

"Another reason for questioning model predictions could be the high nitrogen and phosphorus concentrations that are predicted to generate nuisance blooms of benthic algae: 700 µg TN L^{-1} and 90 µg TP L^{-1} in Unit 3 to prevent pH violations and 1,000 µg TN L^{-1} and 140 µg TP L^{-1} in Unit 4 to prevent nuisance benthic algal problems. Although we know relatively little about nutrient concentrations affecting pH in river, these phosphorus concentrations are many times higher than phosphorus concentrations thought to cause nuisance levels of benthic algal biomass, e.g. greater than 150 mg chl a m^{-2}. Admittedly, there's a great range limiting and saturating nutrient concentrations in the literature, but a 30 µg TP/L benchmark was proposed in the Clark Fork, which

is upstream from this location. Why have higher numbers in the larger mainstem of the Yellowstone River? If we assume Leibig's law of the minimum, and nitrogen and light are sufficiently great to allow algae to grow, why wouldn't the marginal habitats of the Yellowstone River generate nuisance algal biomasses at 30 µg TP/L? At least one reason could explain that discrepancy. The reactive portion of the TP may be lower in the Yellowstone River than in smaller streams where nuisance blooms of benthic algae commonly occur at TP concentrations around 30 µg TP/L. The soluble fractions of total nutrient concentrations, assumed to be the most readily available fractions, were very low in the Yellowstone River during low flow conditions (Table 6-6). However, caution should be exercised when assuming only the soluble fraction of TP is bioavailable; mounting evidence indicates that entrained particulate P and N are recycled in benthic algal mats."

Higher criteria occur in the Yellowstone River for two reasons. First, the response to nutrients is integrated over the wadeable region (<1 m depth), which as Hynes (1969) points out, means that only a portion of the river bottom will be conducive to algal colonization and growth. The second is river turbidity which is considerably higher than western Montana wadeable streams. Hence the comparison between the Yellowstone River and the Clark Fork River by the reviewer is not valid. They are in fact different ecoregions, the lower Yellowstone is significantly more turbid and deeper than the Clark Fork River, and finally the former drains to the Missouri River and the latter to the Columbia River. The reviewer is correct though in one regard, that the Yellowstone River should still grow algal biomasses on the margin of the river at lower nutrient levels; this is the very reason we developed the AT2K model, i.e., to integrate the effect over the entire management area.

With this in mind, the manner in which management endpoints are computed strongly affect the criteria. For example, we used the average benthic algal biomass that develops in the wadeable zone (defined as depths of ≤ 1 m) as our regulatory endpoint. By doing this, it means that algae in the deeper regions of this zone are significantly light limited, and thus the areal-average response is lower. If we managed the river so that no stone were to exceed 150 mg Chl$a$/m^2, the criteria would be different and would be nearer the levels suggested by the reviewer [around 35 µg/L SIN and 10 µg/L SRP which if applied to the soluble regressions of Biggs, (2000) and Dodds et al. (1997) yield biomasses that are less than, or very close to nuisance levels]. However, regulation of a single stone (i.e., the single highest algae level) would not be consistent with the way the algal biomass threshold was derived. For example, the basis of Suplee et al. (2009) was that participants were shown photos of entire river reaches and were asked their impressions (acceptable/non-acceptable) of the entire scene. Since the impressions would be based on the overall appearance of the algae levels (not a single point), and, correspondingly, the algae biomass values provided were the reach averages (of n=10 to 20 replicates), we must regulate biomass for the average of the wadeable region, not the single highest Chl$a$ value recorded (i.e., the single most-green stone).

"The model prediction that low DO is not likely in the Yellowstone River seems reasonable. The Yellowstone River is relatively hydrologically stable, so it is probably not prone to types of extreme low flow events that allow development of low DO with resulting fish kills. Rivers and streams are probably much more susceptible to high pH and fluctuating pH conditions than to low DO; but both phenomena have not been studied sufficiently to understand thoroughly."

We concur with this statement, and also point out that choosing a process based model allowed us to understand both DO and pH dynamics, something that cannot be determined through statistical methods. Thus there is merit to the mechanistic approaches beyond what could be determined using empirical analysis.

4. **Please comment on the appropriateness of using response variables, such as chl-a and pH, as model endpoints for numeric criteria derivation, and thus protection of water quality from nutrient pollution. Please comment on the spatial application of different response variables for deriving numeric nutrient criteria (pH was used for the upstream segment while benthic algal biomass was used in the downstream segment).**

"pH and algal biomass response are appropriate endpoints for justification of nutrient criteria. pH is more directly linked to negative effects on aquatic fauna than nutrient concentrations, so pH is a more proximate threat to a valued ecological attribute. High algal biomass is known be an aesthetic problem in rivers, as established in the great study by Suplee et al. As described above, nutrient criteria above minimally disturbed conditions that prevent nuisance algal accumulations and violation of pH and DO standards may not protect biodiversity of some nutrient-sensitive taxa; however chl a and pH, as well as DO, are appropriate endpoints for protecting designated uses.

The most sensitive response (e.g. chl a, pH, or DO) to a stressor (i.e. nutrients in this case) should be used to establish stressor criteria, even if different response endpoints are the most sensitive in different types of habitats (in this case shallow and deep river habitats). An important goal of environmental management should be protection of ecosystem services. Of course all ecosystem services should not have to be protected in all waters, but appropriate protection is warranted. Montana DEQ and presumably a majority of the people of Montana have supported water quality criteria related to pH and benthic algae. So nutrient concentrations should not be allowed that would generate unacceptable risk of violating the pH and nuisance algal biomass criteria.

The focus on shoreline algal biomass was also appropriate because that is where people most commonly observe the water as they use the resource for recreational purposes."

We agree with this comment.

5. **What other analytical methods would you suggest for deriving numeric nutrient criteria for the mainstem Yellowstone River?**

"The appropriate methods were used to gather information about the development of nutrient criteria, but the results of the computer model were overstated and overweighted in a premature decision on nutrient criteria.

Processed based (computer) models are very informative and valuable, but they are just one line of information. Three basic research approaches can be used to develop numeric nutrient criteria: observing patterns in nature and quantifying relationships between nutrients and key endpoint variables with by statistical models (e.g. regression models); simulating patterns in nature using process-based models; and experiments in controlled environments in which environmental conditions are purposefully manipulated. Each of these methods complement each other. When they all do not agree, then conclusions are suspect. In this case, the predictions of the computer model do not match results of other research based on statistical models and experiments. Even though there are plausible reasons for those discrepancies, there is little reason that the computer model is accurate."

We do **not** believe our results were *"overstated and overweighted"* and defer to reviewer 2's comment in support of this. Also, we factually disagree with the reviewer that it is appropriate to draw direct parallels between the computer simulation and other methods suggested (e.g., statistical

0015242

and experimental). They reflect distinctly different processes, meaning we shouldn't try to force a large river into a wadeable streams approach! A large river (e.g., the Yellowstone) has great spatial variability in light (as described throughout the draft report) whereas wadeable streams are shallow and homogenous. While similar methods could be used to develop statistical or experimental procedures for large rivers, the reviewer misses a very important part of water quality management, that is algal biomass is just one endpoint of consideration. What about pH, DO, or any other important water quality indicators? How would we evaluate their response if just a statistical model of biomass was used? Even if we had such a model, could the model be extended suitably to ascertain criteria? Finally, if streamside mesocosm experiments were completed, would these be comparable to a large river which is primarily deep and turbid and has large underwater areas unsuitable for significant algae colonization? These are questions we asked ourselves prior to initiating the project and simply couldn't answer (even if we combined the statistical and experimental methods). Thus in our opinion modeling is the best line of evidence for criteria determination. Other methods were considered (reference sites, the literature, and experiments) but these were not used due to their inherent limitations in representing the large river response.

"Despite that lack of fit between computer model predictions and measured conditions in the river, during both calibration and validation, the computer model was used. In a simple comparison of accuracy of the computer model predictions of high algal biomass as a result of higher nutrient concentrations (Figure 10-5) and the regression model characterizations between algal biomass and either TN or TP (Figure 15-2), show the regression model warranted more credibility. For the computer model, there was no relationship between algal biomass predicted and the algal biomass observed at stations (Figure 10-5). Plotting these abundances in Figure 10-5 on a log-log scale may have improved the apparent fit, but lack of fit at higher biomasses is likely. Remember the discussion above about lack of data points above 50 mg chl a m^{-2} and poor range of observed conditions. For the regression models, the results were variable but plausible (Figure 15-2). If N:P ratios are low and N limits algal growth, then we'd expect a relationship between algal biomass and TN and not between algal biomass and TP concentration. The range of TP concentrations (and bioavailable P indicated by those concentrations) may have been above the TP concentration considered to have strong effects on benthic algal growth (e.g. 30 μg TP/L). The range of TN concentrations may have crossed the sensitive range and below the limiting nutrient concentration for TN; therefore TN may have been the primary limiting nutrient in the Yellowstone River. Thus, the Montana DEQ got a relationship between TN concentrations and benthic algal biomass, but not TP concentrations and benthic algal biomass. I disagree with the interpretation by Montana DEQ about these relationships. These relationships do show that TN concentrations below 505 μg TN/L should constrain average algal biomass to less than 150 mg chl a m^{-2}, but the lack of significance in the TP algal biomass relationship indicates it should not be used to set a TP criterion. This relationship between TN and algal biomass is really the only evidence in the report for nutrient regulation of benthic algal biomass."

The reviewer suggests that a *"lack of fit"* between the observed and predicted plot (Figure 10-5) makes the mechanistic model unreliable and less suitable than the algal biomass regression in Figure 15-2. However, no evidence is provided supporting this statement. In fact, we have shown previously through analysis of RMSE that the errors are comparable to refereed literature (which is frequently relied on for nutrient criteria). Similarly the use of loose statistical dependence as shown in Figure 15-2 (r^2=0.34, which the reviewer forgot to point out is also log-scale) would be careless given the way the data is collected (oriented toward the shallow regions) and simply not the best available information (which we have compiled via the data collection and modeling). Finally, even

0015243

if the regressions mentioned by the reviewer were suitable, they cannot be extrapolated beyond the observed data, which is a problem give that the concentrations in the river are well below nuisance levels. All of that said, we do agree on one thing, that the TP regression in Figure 15-2 is not useful and we will revise the report indicating this.

"If benthic algal biomass is not simulated accurately by the computer model, can we trust predictions of pH and DO that respond to changes in algal biomass? pH and DO predictions of the computer model were also not validated well because of low sample sizes and ranges of conditions in which the model was calibrated."

We feel this comment is misleading and we have shown that both DO and pH were reliably simulated in two separate August low flow conditions in 2000 and 2007 including both spatial averages and associated diurnal variability. We also stress that the DO and pH response are implicitly a function of the photosynthetic response, which in the case of the Yellowstone River was directly driven by benthic algae. So even if our biomass point measurements did not match perfectly, the community response was correct (as substantiated by reviewer 2's comments). Also, we also point out that the *"low"* sample size mentioned by the reviewer was on par with any academic modeling study (nationally and internationally) and that conditions in the study were sufficiently variable for the intended analysis.

"Another question develops about whether TP concentrations need to be kept below a TP criterion that would constrain algal biomass, if TN concentrations are below that 505 µg/L; but that question is a policy deeper policy question. If TN is kept below 505 µg/L, then presumably there would not be a response of benthic algae to TP if N is the primary limiting nutrient. However, the 505 TN and 30-60 TP range seem close to what I would expect to be saturating nutrient concentrations. So, a combination of TN and TP criteria would provide double protection against risk of high algal biomass."

We agree that criteria levels for both TN and TP are protective and should accommodate future shifts in nutrient availability. That said, water quality managers must use common sense when determining nutrient control strategies and permitted load limits. According to Liebig's law of the minimum, a single available resource (e.g., soluble N or P) will limit yields at a given time which implies that only a single nutrient should be considered in management (unless they are both close to limiting, e.g., co-limiting). Soluble concentrations are difficult to quantify however (Dodds, 2003), and thus we have used the rate of uptake/recycle and associated transport in the model to determine how total nutrients at one point relate to conditions at another (note: these points are different longitudinally because of advection). Given that minimum acceptable nutrient criteria outlined by U.S. EPA were total nutrients, and the fact that totals better lend themselves to ambient nutrient monitoring, permit compliance, and monitoring, we thought this was the most reasonable approach toward criteria development.

"Good calibration of models, computer or regression, should not be expected in a river without a good range of nutrient that result in algal problems at some place across the range of nutrient conditions. In habitats in which no algal problems are observed, it is possible that sediments and low light constrain algal accumulation such that nutrients have no effect on instream algal related conditions. In this case, downstream effects should be the concern/endpoints of criteria. Alternatively, it is possible that most that we know about the asymptotic relationship between nutrient concentrations and algal biomass is not true; or for some other reason, TP concentrations

above 50-100 µg TP/L do regulate benthic algal biomass. Then the high nutrient concentrations as those proposed (700 µg TN L^{-1} and 90 µg TP L^{-1} in Unit 3 to prevent pH violations and 1,000 µg TN L^{-1} and 140 µg TP L^{-1} in Unit 4 to prevent nuisance benthic algal problems) would be appropriate in the Yellowstone River."

We calibrated the models to a range of nutrient conditions so we are not entirely sure where the reviewer is coming from by suggesting the calibration was insufficient (recall soluble N varied longitudinally from 3 to >100 µg/L and variants in soluble P were from ≈3-20 µg/L). Additionally, as mentioned previously, we have shown site-specific environmental considerations (e.g., light) do in fact play a significant role in the productivity of the Yellowstone River (see Figure 2-2 and 2-3 and associated discussion). Lastly we have in fact defined the asymptotic relationship between ambient nutrient levels and biomass response (among other variables) through the model. The response of the Yellowstone River is different than saturating responses of other methods because of, as stated previously, gradients in light. So in fact we are not suggesting, "...*that most that we know about the asymptotic relationship between nutrient concentrations and algal biomass is not true...*" rather that the conditions of previous studies are far different than our application, which is why we chose a modeling approach in the first place. With that in mind, the levels determined in the study are not surprising. In fact, they are very comparable to concentrations suggested for other light-limited wadeable streams in eastern Montana, e.g. ≈1,400 µg/L TN and ≈140 µg/L TP for the Northwestern Great Plains (Suplee and Watson, 2012).

"Continued research in the form of monitoring of the Yellowstone River, surveys of other large rivers, experimental research, and computer modeling will be needed to develop nutrient criteria that protect ecosystem services of large rivers without overprotection. Continued monitoring in the Yellowstone River will enable assessment of whether nutrient concentrations are increasing and nuisance algal biomasses and high pH are becoming more frequent. This will forewarn managers that nutrient related problems are developing and will provide the additional information needed for better computer and regression models used to establish nutrient criteria. In the report, Montana DEQ did propose continued monitoring and data analysis with one goal being learning more about nutrient effects in the river for potential revision of the proposed nutrient criteria. But will reducing the nutrient criteria, based on new science, be practical politically. Why will the public believe the new science if the old science was not sufficient? Why hurry to have nutrient criteria if there are no known problems? Was this the wrong place to try to develop nutrient criteria for large rivers?

A concerted national effort should be developed and maintained to gather the kind of information needed for developing nutrient criteria in large rivers. Monitoring data as well as experimental results should be gathered and evaluated with statistical models and integrated in processed based models to provide sufficient information for development of nutrient criteria in large rivers. Great similarities exist among the large rivers of the world, such that information learned in multiple rivers should be able to be synthesized and related to other large rivers. Until this information is gathered and analyzed, perhaps the most prudent nutrient management strategy is to try to maintain current conditions if there are no existing problems."

We agree with the reviewer that additional computer modeling and further surveying/sampling of the Yellowstone River is important going forward; we point out that Montana has been one of the most active states when it comes to lotic nutrient standards development. Relative to the need for additional study, the reviewer assumes that continued research would result in more stringent criteria for the Yellowstone River, when in fact the standards could go either way. As a matter of point, the

standards would also need to be changed if the beneficial uses of the river currently in law were to be changed by the public. The political reality of water quality standards is they are updated constantly which is why the Clean Water Act requires states to review them every 3 years. Sometimes standards are made more stringent, sometimes relaxed. Our experience in this matter has been that the public accepts improving engineering/science, and that these advances can result in changed water quality standards.

We disagree with the reviewer's suggestion that this may be the wrong river to study at the wrong time. Water quality standards are not just for polluted rivers they are also to protect those that are still healthy; that's why all states have non-degradation policies as part of their water quality standards. The Yellowstone is one of the fastest growing regions of the state (e.g., Billings population increased about 15% from 2000-2010) and nutrient-laden discharges from urban areas will only steadily increase. We selected this river segment because it was un-impounded (which simplifies modeling and interpretation of applicable water quality laws), it is well gaged, and reasonably reported on in both the open- and grey-literature. This means that there was a good chance of successfully developing nutrient criteria for the river.

Finally, we agree with the reviewer that a concerted national effort to gather data on large rivers including the use of modeling and experimental research would be valuable. We hope that the academic community will undertake such work. However, our finding has been that national efforts to develop numeric nutrient standards for large rivers by anyone, academic, governmental, or private, has been slim to none. This has occurred is in spite of the fact that former Vice President Gore's Clean Water Action Plan, which called on states to develop numeric nutrient criteria for waterbodies, was published in the Federal Register in 1998, fourteen years ago! Work on large river nutrient standards needs to be started by someone, somewhere, and we feel our study was an excellent start. We believe the use of existing water quality models (and development of new models such as the one described by DEQ) will help advance criteria-development methods nationally. Note that the Water Environment Research Foundation (WERF) is currently researching the use of such models for site-specific criteria determination.

"A couple editorial changes worthy of note:
Figure 9-1 makes much more sense to me if Table 8-1 were changed to Table 9-1."

Thank you. The section numbering changed several times and we did not get corrected in Figure 9-1. We will make this change.

"Figures 13-4 and 15-2 were hard to understand because the independent variable (nutrient concentration) was not on the X axis."

We have received this comment from reviewer 2 and will make this change. We had initially plotted the state-variable of interest on the abscissa and the criteria on the ordinate as nutrient criteria are really the dependent variable. However, since this apparently has been confusing to a number of people, we will make the change.

SECTION 2.0 - Responses to Reviewer 2

"Using a Computer Water Quality Model to Derive Numeric Nutrient Criteria, Lower Yellowstone River, MT (Montana DEQ, 2011) provides a comprehensive discussion of Montana Department of Environmental Quality (DEQ) efforts to develop nitrogen and phosphorus criteria for the lower Yellowstone. This is done through the development of a site-specific mechanistic nutrient response model that links nutrient loads to measurable endpoints associated with the support of designated uses in the river. The approach is consistent with EPA guidance on establishing TMDLs to address narrative nutrient criteria, which also results in site-specific objectives.

The result of the study is recommendations on site-specific nutrient criteria for the Lower Yellowstone. The results are truly site-specific as they depend on the conditions present in the Lower Yellowstone and it is not clear that they would be applicable to other, similar waterbodies. The results could serve as a template for the derivation of site-specific criteria for other large rivers; however, the evidently high level of effort required to complete this study may preclude wide application.

In general, the modeling and analysis presented here is well done and adequately documented. There are, however, some specific questions that should be resolved before finalizing the analysis. These are described below.

The site-specific nutrient response approach is attractive for several reasons. As noted by DEQ, there is a lack of reference watersheds for large rivers, and methods appropriate to wadeable streams are not transferable to large rivers. In addition, nutrients themselves (except at extreme concentrations) generally do not directly impair designated uses; instead, it is the secondary effects of elevated nutrients, generally involving algal growth, that lead to use impairment. These secondary effects differ according to site characteristics, such as light availability, residence time, and scour regime, which means that the assimilative capacity of a waterbody for nutrients is inherently site-specific and determined by a variety of co-factors; thus the most economically efficient nutrient criteria should also be site-specific.

DEQ has developed site-specific criteria for the lower Yellowstone that reflect specific characteristics of the basin. Notably, the river is deep and turbid, both of which characteristics reduce light availability and thus also reduce the expression of nutrient impacts through algal growth. In other words, these characteristics of the Yellowstone River serve to increase its assimilative capacity for nutrients.

It is clearly appropriate to consider the hydrologic characteristics of the Yellowstone in developing site specific criteria. In particular, the amount of flow and depth of the river, which reduce the area in which benthic algae can grow, is a largely natural condition. The case for turbidity is a little less clear. The tributaries of the Yellowstone, especially the Powder River, are believed to be naturally turbid. However, the present day turbidity is also affected by land use practices (silviculture, agriculture, grazing, mineral extraction, etc.). If turbidity is greatly elevated by anthropogenic sources then it would appear inappropriate to count the full effect of high turbidity on reducing algal growth as a "credit" that allows for higher nutrient concentrations.

The report (p. 4-8) says, regarding sediment loads in the Powder River, "Much of its contribution may be natural. A number of other anthropogenic non-point sources are believed to occur..." There are turbidity standards for the lower Yellowstone. These allow a maximum increase of 10 NTU relative to natural conditions (Table 4-3). The lower Yellowstone has not been assessed as impaired by turbidity, but it is not clear if an analysis of natural turbidity levels in the system has been performed. It would appear most appropriate to evaluate nutrient criteria with turbidity constrained to meet standards – i.e.,

the natural turbidity regime plus 10 NTU. At a minimum, the report should discuss these issues and make a case for the selected approach."

We had debated the reviewer's consideration prior to the publishing of our draft report and came to the conclusion that a large percentage of the sediment load in the river was natural. We did not state why however. Our justification is as follows. First, a fairly large increase in turbidity occurred downstream of the Powder River when in fact there was no flow contribution to account for such changes. Peterson and Porter (2002) note similar findings writing, *"Water turbidity increased two-fold between Y7 (Forsyth) and Y8 (Miles City), downstream from the Bighorn and Tongue River tributary confluences, then increased from 12 NTU at Y9 (Terry) to 24 NTU at Y11 (Glendive), downstream from the Powder River confluence. However, the Powder River was dry prior to and during the time of sampling in late August 2000"*. Given that both studies found similar changes at similar times (i.e., when the Powder River had very little flow), we concluded that a large unaccounted for autochthonous source exists in the lower river. Most likely it is previously deposited sediment from the Powder River. Still it is unclear whether this load is natural or anthropogenic.

The historical description below provides a persuasive argument clarifying DEQ's argument for natural. Vance et al. (2006) indicate that Francois Antoine Laroque, passed through the lower Yellowstone in the early 1800's (prior to Lewis and Clark). He describes, *"The Powder River is here about ¾ acre in breadth, its water middling deep, but it appears to have risen lately as a quantity of leaves and wood was drifting on it…It is amazing how very barren the ground is between this and the less Missouri, nothing can hardly be seen but those Corne de Racquettes (prickly pear cactus). Our horses are nearly starved. There is grass in the woods but none in the plains…The current of the river is very strong and the water so muddy that it is hardly drinkable. The savages say that it is always thus and that is the reason that they call it Powder River; from the quantity of drifting fine sand set in motion by the coast wind which blinds people and dirtys the water."*

Similarly, on Friday July 30th, 1806, William Clark of the Lewis and Clark expedition noted, *"Here is the first appearance of Birnt hills which I have Seen on this river they are at a distance from the river on the Lard Side…after the rain and wind passed over I proceeded on at 7 Miles passed the enterance of a river the water of which is 100 yds wide, the bead of this river nearly ¼ of a mile this river is Shallow and the water very muddy and of the Colour of the banks a darkish brown. I observe great quantities of red Stone thrown out of this river that from the appearance of the hills at a distance on its lower Side induced me to call this red Stone river. [NB: By a coincidence I found the Indian name Wa ha Sah] as the water was disagreeably muddy I could not Camp on that Side below its mouth."*

The previous descriptions in our opinion provide convincing evidence that much of the sediment load from the Powder River is natural. After all, it is hard to imagine anthropogenic sources could elevate turbidity above pre-settlement levels by any meaningful amount. To put the magnitude of the load into perspective, NRCS (2009) estimates the current sediment load of the Yellowstone River at Forsyth, MT as being 3,769 ac-ft/yr whereas the Powder River itself has a load of 3,400 ac-ft/yr (nearly the same amount as the entire upper Yellowstone drainage area). So while no formal sediment source assessment exists to quantify the natural and anthropogenic fractions, we feel it is reasonable to conclude that there has always has been a very large natural sediment loading originating from this region and any turbidity that exists during low-flow conditions is likely natural.

"One additional caution regarding the study in general is that the authors take some liberties in reinterpreting numeric criteria from the Administrative Rules of Montana into "more appropriate" forms.

- *Total dissolved gas levels must be ≤ 110 percent of saturation: The Montana administrative code seems to establish a clear limit of 110 percent of saturation. The authors argue (p. 13-15) "the standard is mainly intended to control super-saturation of atmospheric gas below dam spillways... A thorough literature review... shows that fish are tolerant of much higher total gas levels than the state's standard when the gas pressure is driven by oxygen. For example, fish have been found to tolerate DO saturation levels to 300% DO without manifesting [gas bubble] disease... DO supersaturation levels observed in our model runs were never greater than 175% of saturation and were therefore not an endpoint of consideration with respect to gas bubble disease..." In my opinion, this argument is sensible; however, it is not what the rule says. Presumably, a modification to the criterion should be needed to eliminate consideration of meeting the dissolved gas target from the analysis.*

- *Induced variation in pH must be less than 0.5 pH units within the range of 6.5 to 9.0 or without change outside this range: This requirement is also established in the Montana code. The authors (p. 13-12) contend that this is mistaken and should reflect a two-part test (greater than 9 units and induced variation of 0.5) "as pH in the range of 6.5-9.0 is considered harmless to fish and diurnal changes (delta) greater than 0.5 are only unacceptable when they push the pH outside the 6.5-9.0 range." As with total dissolved gas, this argument makes some sense, but appears to be at odds with existing regulations."*

The reviewer is correct that we did not adhere strictly to the letter of Montana law in identifying and applying water quality endpoints in the model. Rather, we applied current science relative to the effects of TDG and pH. In doing so we understand that we may expose ourselves to some criticism. However, we felt (especially the author who is in DEQ Water Quality Standards) that scientifically-based model endpoints are more important than upholding an antiquated standard given the real intent of water quality criteria is to protect the uses. That said, water quality criteria are often updated/changed to reflect the current state of the science with the underlying intent always remaining unchanged; that is the protection of fish and aquatic life. The relative shortcomings of the two currently-adopted criteria in question (e.g., the fact that aquatic life tolerate higher TDG if it is DO driven, and the two-part pH test) will likely be addressed by DEQ in a future triennial review. Thus, it was better to use appropriate TDG and pH endpoints with the anticipation that current criteria are likely to be updated anyway.

1. Please evaluate the sufficiency and appropriateness of the data used to run the model.

"An extensive data collection effort was undertaken to support the modeling. This effort was specifically designed to support QUAL2K application. The data included two 10-day synoptic surveys (August and September 2007) at multiple sites, along with YSI sonde deployment at 20 or so mainstem and tributary sites throughout the summer. All water quality data were collected in accordance with a QAPP. In addition, a variety of historical data were also located and documented, including a synoptic USGS data set from August 2000. Data were also collected via algal growth rate experiments.

Three good synoptic data sets should be sufficient to test, calibrate, and validate and steady-state model such as QUAL2K. Additional inputs, such as climate forcing, are well documented

0015249

Estimates of reaeration and SOD are key inputs to QUAL2K modeling and often difficult to disentangle. DEQ used the approximate delta method of McBride and Chapra to estimate reaeration rates from continuous sonde data. The resulting estimates of k_a have 95% confidence limits on the order of about 1-1.5 day^{-1} on mean values from 2-7 day^{-1}. An attempt was made to estimate SOD with in situ chambers, which is the preferred method, but these failed due to the coarse nature of the river substrate that prevented a good seal. Therefore, estimates were instead estimated from incubated cores, resulting in values that are consistent with literature values for sand bottoms (around 0.5 g/m2/d). However, the authors then state that "percent SOD coverage was visually estimated at each field transect", resulting in values of either zero or 5 percent "cover by SOD" by reach. This percent cover operates as a scaling factor on SOD; thus the authors have effectively reduced SOD in the model to near zero. How it is possible to determine SOD cover visually is not explained, as the levels cited are typical of sands, not mucks. It further seems unreasonable that reaches can have 80 – 100 percent cover by algae but zero "cover" by SOD. Thus, SOD may be underestimated in the model. This in turn may introduce some bias into the benthic algae and diurnal DO calibration."

The reviewer is correct that SOD was low in the river but these values originated from actual observations (even though they were cores) and deviation from them would simply not be justified. In characterizing the percentage of the river that was SOD generating, we relied primarily on the substrate characterization in the field. We observed sediment at 11 locations within each sampling transect and used particle size (i.e., fine grained) as a surrogate for SOD generating material (which were characteristic of our core measurements). In all cases, <5% of the channel substrate would qualify as SOD responsive, which is shown in Table 8-7. In many instances none of the sampling locations contained fines. Admittedly, our *n* was small, but observations did generally fit our conceptual understanding of the river, that is it comprises a well-armored cobble/gravel bed with high flow velocities devoid of organics or other SOD generating material. This does not mean that depositional areas/shallow-water environments where higher SOD (mucks) are not present. Review of aerial photography indicates that such areas do exist, primarily behind the Cartersville diversion dam near Forsyth and in braids and oxbows. The overall spatial extent of these areas is small relative to the channel however. Finally, as the reviewer is aware of, SOD is a direct scaling factor on the oxygen mass balance. Assuming respiration, reaeration, and nitrification are reasonably known (which they were), the leftover deficit (which was small) would have to be attributed to SOD.

In regard to the algal cover percentage (80-100%), again we relied on field data. While percent cover is again a subjective measurement, we find no reason to deviate from our original observations. Admittedly, the water was too deep to make a visual assessment in several instances (noted as <u>not</u> visible on the field form), but the presence Chl*a* was verified analytically at nearly all transect sites (even on sands/clays). Lastly, the percentages applied in the model are consistent with diurnal oxygen (DO) profiles of the river. For example, in order to meet the productivity response of the river, an areal coverage of 100% was required.

"Another potential area where data are somewhat weak is in the estimates of groundwater quality. This input is based on wells less than 200 feet deep and within 5 km of the river. The problem is the assumption that well measurements are equivalent to the quality of water that discharges from groundwater to the river. Typically there can be significant amounts of nutrient uptake by sediment bacteria during the seepage process. This, however, appears to constitute only a very small portion of the total nutrient mass balance and so is not a significant cause for concern."

0015250

We concur that the groundwater contribution is hard to estimate due to the reasons mentioned by the reviewer. We also suggest though that this is not a major concern due to the following reasons: (1) flow at the upstream boundary encompasses nearly 70% of the inflow to the study reach and groundwater flux comprised less than half of the remaining 30%, (2) estimated groundwater loads (Figure 17-5) were only 4.6% and 1.8% of the soluble N and P supply to the river, and (3) groundwater nutrient concentrations were not considerably high relative to polluted aquifers. So while the values used in the model could be in error due to the reasons mentioned by the reviewer, we feel this would likely result in only a small loading error.

2. Please evaluate the model calibration and validation

"Calibration was performed on the August 2007 dataset with validation on the September 2007 dataset. An additional validation test was undertaken with 2000 USGS data. The calibration was carried out in accordance with a plan and criteria pre-specified in the QAPP for temperature, DO, phytoplankton chlorophyll a, and bottom algae chlorophyll a. The authors are commended for using the approach of pre-specifying criteria, which is consistent with EPA QA recommendations, but often not done in modeling studies. One concern with the approach is that the QAPP criteria are not based on an analysis of the level of precision needed to meet decision needs under a systematic planning approach but rather seem to be mostly derived from literature recommendations. (The QAPP does not actually state the basis for the selection of the criteria). The specified criteria for Relative Error and Root Mean Squared Error are aggressive but feasible for temperature (±5% or 1° C) and dissolved oxygen (±10% or 0.5 mg/L). The targets for chlorophyll a (±10% for phytoplankton and ±20% for bottom algae) are, in my experience, more stringent than is likely to be attainable for models of this type – particularly for bottom algae chlorophyll a, as this is affected by a variety of processes, including grazing, scour, and variability in the carbon:chlorophyll a ratio, that make precise prediction difficult. The QAPP did not specify acceptance criteria for the pH calibration, as pH was not identified as an important decision variable until after development of the model. It would also have been desirable to specify acceptance criteria for the nutrient simulation (e.g., ±25%), but it would not be appropriate to add acceptance criteria after the fact."

The reviewer is correct that we probably did not do enough up-front consideration of model acceptance criteria (i.e. based on the level of precision needed to meet decision needs) but rather relied on the literature. However, the primary reason was that prior precedent does not exist for making these decisions. For example it was unclear (at least to us) what level of precision may be needed to make acceptable decisions (e.g., would the system be very sensitive to nutrient additions, how does the pH respond, etc.). We would have for the most part been relying on professional judgment. We are also in agreement with the reviewer that our state-variable targets were probably too aggressive. In hindsight, it would have been nice to have provided more flexibility in these values, as well as specifying pH and total nutrient targets *a priori*. In this regard, we will now have to work through these considerations in development of the criteria using knowledge about uncertainty and past criteria development efforts.

"Model parameters and rate coefficients adjusted during calibration are clearly documented and compared to literature values – in most cases. For some reason, the literature ranges for algal stoichiometry and various Arrhenius temperature coefficients are cited as "n/a", although citations are available; however, none of these values look to be unreasonable."

The reviewer is correct. Stoichiometry values can be found in Bowie et al. (1985) and have also been recommended by Chapra et al. (2008). We will revise the table to include suggested ranges. In regard to the temperature coefficients, we did not calibrate these values and we will note that in the footnote of the table.

"Results of model calibration and validation (both September 2007 and August 2000) are summarized in Table 10-1, where it is stated that the QAPP criteria are met except for benthic algae. This is not quite correct, as the Relative Error for DO in the 2nd validation is 18.5%, greater than the criterion of ±10%."

Thank you for finding this mistake. We will revise the table and text.

"Most aspects of the model fit appear quite good. One problem area is the nitrogen simulation. While total N is fit well, there are large relative errors in the nitrate and ammonium simulations. The model consistently underestimates observed NH_4-N concentrations, while overestimating NO_2+NO_3-N during the calibration and underestimating it during the validation. The authors suggest that this is mostly due to changes in trophic condition between August and September, but it looks as though there is something else occurring, probably associated with estimated boundary conditions for incremental inflows."

We agree with the first part of this comment and will investigate how minor recalibration to reduce the nitrification rate will influence the simulation (thereby increasing NH_4-N and decreasing NO_2+NO_3-N). We expect that such a change will probably have a greater effect on ammonium than nitrate/nitrite given their comparative concentrations. Relative to the change in trophic condition, we still contend that shift in river photosynthetic response is the most valid hypothesis, more so than the shift in incremental flows and associated boundary conditions as suggested by the reviewer. For example, we made it a point to evaluate different flow and concentration conditions for each period (August and September) for both tributaries and irrigation return flows as described in Section 7. While some of this data was regressed/estimated, it was reasonably similar both months. Likewise, the relative contribution of these sources was small in comparison to the overall headwater boundary condition soluble nutrient load (as previously noted, referring to the fact that the headwater constituted 70% of the available nutrient load to the reach). In our opinion then, the magnitude of such errors would not be sufficient to cause the large difference observed between the two periods. Autotrophic response just simply slowed (e.g., nutrient uptake, diurnal variation in DO and pH, etc.) which combined with other indicators (i.e., algal physiology evaluations, water temperature, daylength, etc.) make us believe the change in photosynthetic response and resulting nutrient uptake was driven by algal senescence.

"In addition to the base QUAL2K model, the authors made use of several related tools. First, they worked cooperatively with Tufts University to develop a new model, AlgaeTransect2K (AT2K) that relates longitudinal QUAL2K model output to lateral benthic algal density. This tool was designed to account for lateral heterogeneity in areas where only the wadeable, nearshore areas have sufficient light to support significant bottom algae growth. It is not entirely clear how well AT2K works when applied essentially as a post-processor to QUAL2K. That is, the QUAL2K model calibration relies on laterally averaged conditions – including the effects of benthic algal growth calculated based on mean depth. As the relationship between depth and light attenuation is not linear it would not seem appropriate to apply AT2K as a post-processor to QUAL2K results; rather the laterally averaged

0015252

bottom algae density from AT2K would seem to need to be re-input to QUAL2K in an iterative process until convergence was obtained."

We do not know a good way to characterize the utility of AT2K when applied as a post-processor to QUAL2K other than to suggest the following: (1) simulated areal biomasses when laterally averaged are nearly identical to the lateral average in QUAL2K (meaning both models converge on the same areal biomass) and (2) calibration of both models was done with only a single set of rate coefficients so that the kinetics in each model are identical despite their difference in conceptual representation. That said, the problem described by the reviewer is plausible and illustrates at least one potential concern when dealing with multi-dimensional water quality problems. Transect station-specific computations from AT2K could in fact be theoretically differ from laterally averaged computations in Q2K, especially with regard to spatial differences in river productivity. These differences would be most likely to affect the oxygen and pH mass balances but it seems like the spatial errors cancel otherwise depth- and width- averaged results from the longitudinal model would not be correct. Thus the calibration method employed by DEQ (i.e., adjustment of rates in both models until acceptable agreement in both models was achieved) seems like the most reasonable method and is valid for discerning the spatial detail of periphyton at a given river transect (instead of transfer of forcing or biomass data as suggested by the reviewer).

"The apparently weak fit to observed benthic algae chlorophyll a is of less concern, as this measure is typically highly variable both in space and time. The fact that both the longitudinal and diurnal profiles of DO and pH are well simulated suggests that the algal simulation is acceptable."

We wholeheartedly agree and have made it a point to stress this as part of our response to reviewer 1 (who has a different opinion). Diurnal DO and pH give the true integrated effect of algal community processes which are equally, or perhaps more important, than noisy point algal measurements.

"Several additional minor criticisms regarding the calibration are:

- *The groundwater contribution was treated as the only unknown in the flow balance (p. 7-9). In fact, irrigation lateral return flows are entirely estimated, although a regression relationship is cited. This uncertainty in the estimate of groundwater accrual should be noted."*

We will revise the text on 7-9 to make this more apparent. We had put some text on page 7-8 describing this, and had a footnote on page 7-33, but we will revise the groundwater discussion on 7-9 so it isn't perceived as misleading.

- *"Evaporation losses from the river are modeled as diffuse abstractions, which remove constituent mass as well. DEQ recognized this as an issue, but the model has not yet been modified to allow removal of water only."*

A beta version of Q2K now has this functionality, but at this point it is not practical to apply the new version of the model given the significant effort to reconfigure the report and associated modeling results (even though very little change is expected). Given that evaporation is a very small portion of the water balance (see page 7-9), we feel it is OK to proceed as currently proposed.

0015253

3. Please evaluate the model calibration and validation

"Uncertainty in model predictions, as shown by the calibration and validation exercises, is fully acknowledged and discussed in some detail in the text. In addition, Chapter 14 presents an error propagation analysis in which the effect of uncertainty in boundary conditions, model parameters, and rate coefficients on model predictions is examined. This was accomplished through Monte Carlo analysis using QUAL2K-UNCAS, a re-write of the original QUAL2E-UNCAS uncertainty analysis. (This model version does not appear to be publicly available.). Headwater boundary conditions appear to be the most sensitive parameter controlling pH (which is significant, as pH becomes the decision criterion for the upper reach). However, this conclusion would be better supported if sensitivity to irrigation return flows was also evaluated."

The reviewer is correct that UNCAS for QUAL2K is not in the public domain and awaits publication. Contrary to as suggested by the reviewer though, we <u>did</u> evaluate irrigation return flows as part of the UNCAS work in Section 8.0. Confusion about this may result from the fact that the nomenclature of the analysis was not clear. Large irrigation canals were included in the "point source" evaluation whereas lateral return flows were included in the "diffuse source" component. Another thing that may have added to the confusion is that NSC values for these boundary conditions were not in Table 8-1 and 8-2 (because DO, pH, benthic algae, and TN/TP were highly insensitive to their changes). We will add some text in both Sections 8 and Section 14 clarifying this.

"The major problem with the uncertainty analysis is the interpretation of results. These focus on the variance in output for TN and TP as a function of input uncertainty (excluding nutrient loads), which are used to suggest that the confidence limits on the proposed criteria are small. This approach is incomplete. Instead of TN and TP, the authors should be examining the effect of error propagation on response variables used to derive the criteria. For example, if the error propagation analysis resulted in large confidence limits in predicted benthic algal density it would be appropriate to set lower nutrient criteria to account for this uncertainty."

We think this is a perceptive comment and an oversight on our part. We will re-examine the perturbation variance of ecological responses (i.e., by including pH and benthic algae) as part of the final report. We will then use this information to draw better conclusions, if necessary.

4. Please comment on the appropriateness of using response variables, such as chl-a and pH, as model endpoints for numeric criteria derivation, and thus protection of water quality from nutrient pollution. Please comment on the spatial application of different response variables for deriving numeric nutrient criteria (pH was used for the upstream segment while benthic algal biomass was used in the downstream segment).

"The approach of using response variables is wholly appropriate for establishing site-specific nutrient criteria. The response variable analysis (if comprehensive) ensures that factors that actually impair designated uses are controlled to acceptable levels as a result of nutrient limits while protecting against the economic impacts of unnecessarily stringent limits based on generic nutrient concentration objectives. It is important, however, to ensure that all secondary impacts of nutrient concentrations that have a potential to impair uses are considered in this type of approach.

The response variable approach appropriately relies on the most limiting response in each reach. That is, each response variable must be controlled within criterion concentrations and other

appropriate limits. pH is the most limiting response in the upstream segment and benthic algal biomass the most limiting response in the downstream segment; however, the proposed criteria will protect both pH and benthic algal biomass in all analyzed segments of the river. Thus, the approach is appropriate.

Application of the model was conducted using 14Q10 flows, typical August meteorology, and low-flow tributary boundary conditions. Selection of these conditions is well supported and documented in Chapter 12.

The model predicts that there is additional assimilative capacity for nutrients under current conditions. Therefore, the model was used to evaluate nutrient criteria by simulating nutrient additions of NO3 or soluble reactive P (SRP) that achieve new concentration levels in stream – requiring an iterative procedure. Ten levels of NO3 (with SRP at non-limiting levels) and ten levels of SRP (with NO3 at nonlimiting levels) were tested. Resulting TN and TP concentrations were calculated by the model. Output from each test was compared to nutrient-related criteria or recommendations for DO, pH, benthic algal biomass, total dissolved gas, and TOC. Of these, the benthic algal biomass and TOC targets are recommendations, not standards.

The benthic algal biomass target of 150 mg/m2 chlorophyll a (as an average for the wadeable region) is DEQ's recommendation to protect recreational uses. This is certainly relevant to use support; however, some justification should be provided as to whether 150 mg/m2 as a wadeable zone average is adequate to support aquatic life uses as well as recreational uses – especially in light of recommendations for the Clark Fork of 100 mg/m2 as an average and 150 mg/m2 as a maximum density."

A lower benthic algae standard for the Clark Fork River (100 mg Chla/m^2 as a summer average) was recommended along with a 150 mg Chla/m^2 maximum in the 1990s as part of the Voluntary Nutrient Reduction Program (VNRP). However, estimates at this time were based on limited academic literature, which did not include evaluation of the public's opinion on the matter. Subsequently, Suplee et al. (2009) show that the public majority in the Clark Fork basin (i.e., Missoula) are accepting of average algae levels up to 150 mg Chla/m^2 (but no higher). Thus, we believe that the 150 mg Chla/m^2 benchmark is, on average, appropriate. In regard to aquatic life uses, nutrient criteria are determined according to the most sensitive use. So if aquatic life standards were exceeded according to the model (e.g., pH or DO) they were used in establishing the criteria. We do not think that 150 mg Chla/m^2 impairs aquatic life uses in large rivers whereas it does in wadeable streams due to accrual of decomposing algae in pools (resulting DO minima <5 mg/L).

"TOC was compared to EPA recommendations for treatment thresholds to minimize harmful disinfection byproducts, and a footnote states "primarily we are concerned with whether or not any scenario would push the river over a required treatment threshold…", thus requiring a higher level of TOC removal. While this is related to drinking water uses, it appears to be more of an economic than a use-protection argument. The issue is moot, however, as TOC was not a limiting factor in the determination of assimilative capacity.

As mentioned in my introductory remarks, there are some issues with how the authors have interpreted (or re-interpreted) existing Montana water quality standards for pH and dissolved gas. The dissolved gas criterion would exceed the 110 percent threshold defined in the rule, if it was deemed applicable, and might thus require more stringent nutrient limits; however, the authors argue that this is not appropriate. It is stated (p.13-16) that the nutrient addition runs resulted in dissolved gas concentrations up to 175 percent of saturation; however, full details are not provided.

0015255

Regarding pH, this becomes a limiting factor for nutrients primarily because the natural pH of the system seems to be high (> 8.5 at the headwater reach for this analysis); thus only a small increment is needed to push it over the level of 9 standard units. The authors should likely discuss whether there are other anthropogenic causes contributing to elevated pH in the system."

We have already addressed both the total dissolved gas and pH standard interpretation issue earlier in our response (in the introductory remarks). With regard to human-caused factors that may have already elevated the pH of the Yellowstone River, our understanding is that a pH of 8.5 at Forsyth is natural or close to a natural. For example, multi-year monitoring studies show a longitudinal change in pH along the Yellowstone River, from just outside of Yellowstone National Park (median: 7.95) to Livingston (median: 8.0) to Billings (median: 8.2) to Forsyth (median: 8.4) (USGS, 2004). As the reviewer is aware, pH in freshwaters is largely controlled by the carbonate-bicarbonate buffer system (Morel and Hering, 1993) and surface waters in Montana are very often alkaline. Downstream of Billings cretaceous sedimentary rocks underlay the river and contribute to increasing calcium carbonate concentrations that elevate pH (USGS, 2004). In fact, according to the 25th percentile bicarbonate concentration at Forsyth (90 mg/L; USGS, 2004) and open carbonate equilibrium theory (i.e., $H_2CO_3^*=10^{-5}$ molar and $pK_{a1}=6.35$), pH should naturally be approximately 8.5 assuming all bicarbonate is geochemically derived (which seems reasonable using the 25th percentile). Finally, the Big Horn River (upstream of the modeled reach) contributes a large proportion of flow to the Yellowstone River and has a median alkalinity of 188 mg/L as $CaCO_3$ (much higher than the Yellowstone River at Livingston, where median alkalinity is 54 mg/L as $CaCO_3$). The Bighorn basin is dominated by rangeland land uses which for the most part are natural. Thus while we cannot say with 100% absolute certainty that pH in our modeled reach is natural, the suggested values are fairly typical for larger rivers and streams in the Yellowstone River basin (median range: 8.1 to 8.5) (Lambing and Cleasby, 2006) and reasonably approximate natural.

5. **What other analytical methods would you suggest for deriving numeric nutrient criteria for the mainstem Yellowstone River?**

"In my opinion, the approach used is the appropriate one for the lower Yellowstone River as it provides a fairly comprehensive evaluation of stressor-response relationships specific to the site. A variety of other methods could also have been attempted. Most of these are summarized in Chapter 15 and would generally result in lower criteria. This is expected because (except for the continuous modeling option) they do not fully account for (or wholly ignore) the site-specific characteristics of the Yellowstone.
Briefly:

- *Literature provides a wide range of potential nutrient criteria values, some lower and some higher than the proposed lower Yellowstone criteria. None of the identified literature sources is fully applicable to a deep, turbid river in the High Plains. General recommendations (such as Dodds, 1997, guidance of 350 µg/L TN and 30 µg/L TP to keep benthic biomass below 150 mg/m2 chlorophyll a can be regarded as a lower bound that might apply if other mitigating factors (turbidity, depth) were not present.*

We agree with this comment and suggest it be referenced to counter reviewer 1's critical review.

- *Reference site approaches are in theory applicable; however, an appropriate unimpacted reference for the Yellowstone does not seem to be available. Setting criteria to an unimpacted reference*

condition would also tend to establish a lower bound level of no anthropogenic effect and not a site-specific estimate of assimilative capacity.

We initially considered a reference site approach (see the QAPP for further detail) but found that the least impacted location was well upstream of the study reach almost entirely in the Middle Rockies ecoregion. Due to the fact that the site had significantly different character than the reach in question (predominantly because of natural reasons), use of the site was omitted.

- *Level III Ecoregional Criteria recommendations are, in essence, a formal summary of available reference site data. These recommendations are most applicable to wadeable streams and do not take conditions specific to the Yellowstone into account.*

- *Regression analysis is presented by DEQ relating benthic algal chlorophyll a to TN and TP in the Yellowstone. This implicitly takes into account some of the site-specific conditions present in the river. These regressions could be used to predict concentrations at which nuisance levels are exceeded; however, the coefficients of determination are quite low, indicating weak predictive ability. Thus the approach of using a calibrated, mechanistic model is preferable. I do suggest that the authors present a multiple regression analysis of benthic algae as a function of both TN and TP, similar to the equations developed by Dodds on the Clark Fork."*

As suggested by the reviewer, we will include a multiple regression analysis (with adjusted r^2) in our final report.

- *"Continuous simulation modeling could also be used to provide a more detailed analysis of nutrient and algal dynamics over time in the Yellowstone. This would primarily be of academic interest, as the identification and simulation of critical conditions using the steady state QUAL2K model appears adequate for the purposes of establishing criteria."*

We agree that a time-variable analysis might be of interest but we will not be pursuing such work given the limited benefit and added complexity. It should be noted that Washington Department of Ecology has just released a beta version of QUAL2Kw with dynamic capability (code from WASP) so this may be a consideration in the future (or for retrospective analysis of the Yellowstone River). Other researchers, i.e., the Water Environment Research Foundation (WERF) are also developing a numeric nutrient criteria toolbox as part of the Link1T11 research proposal (Limnotech, Tufts University, Brown and Caldwell, and others) which will further shed light on such approaches.

SECTION 3.0 - References

Baly, E.C.C. 1935. The Kinetics of Photosynthesis. Proc. Royal Soc. London Ser. B, 117:218-239.

Barnwell, T. O., L. C. Brown, and R. C. Whittemore. 2004. Importance of Field Data in Stream Water Quality Modeling Using QUAL2E-UNCAS. J. Environ. Eng.-ASCE. 130(6): 643-647.

Biggs, B.J.F. 2000. Eutrophication of streams and rivers: Dissolved nutrient-chlorophyll relationships for benthic algae. J. N. Am. Benthol. Soc. 19: 17-31.

Borchardt, M. A. 1996. "Nutrients," in Algal ecology-Freshwater benthic ecosystems, Stevenson, R. Jan, Bothwell, M. L., and Lowe, R. L., (San Diego, CA: Academic Press): 183-227.

Bothwell, M. L. 1985. Phosphorus limitation of lotic periphyton growth rates: An intersite comparison using continuous-flow troughs (Thompson River system, British Columbia). Limnol. Oceanogr. 30(3): 527-542.

Bothwell, M. L. 1989. Phosphorus-limited growth dynamics of lotic periphytic diatom communities: Areal biomass and cellular growth rate responses. Can. J. Fish. Aquat. Sci. 46: 1293-1301.

Bowie, G. L., W. B. Mills, D. B. Porcella, C. L. Campbell, J. R. Pagenkopf, G. L. Rupp, K. M. Johnson, P. W. H. Chan, S. Gherini, and C. E. Chamberlin. 1985. Rates, constants, and kinetics formulations in surface water quality modeling (Second Edition). Athens, GA: United States Environmental Protection Agency. Report EPA/600/3-85/040.

Canham, C.D., Cole, J.J., and W.K. Lauenroth. 2003. Models in ecosystem science. Princeton, NJ: Princeton University Press.

Chapra, S.C. 2003. Engineering water quality models and TMDLs. J. Water Res. Pl.-ASCE. 129(4): 247-256.

Chapra, S.C. 2011. Rubbish, stink, and death: The historical evolution, present state, and future direction of water-quality management and modeling. Environ. Eng. Res. 16(3): 113-119.

Chapra, S. C., G. J. Pelletier, and H. Tao. 2008. A modeling framework for simulating river and stream water quality, Version 2.1: Documentation and users manual. Medford, MA: Civil and Environmental Engineering Department, Tufts University.

Chételat, J., Pick, F.R., and A. Morin. 1999. Periphyton biomass and community composition in rivers of different nutrient status. Can. J. Fish. Aquat. Sci. 56: 560-569.

Di Toro. 1980. Applicability of cellular equilibrium and Monod theory to phytoplankton growth kinetics. Ecol. Model. 8: 201-218.

Dodds, W.K., 1991. Community interactions between the filamentous algal *Cladophora glomerata (L.) Kuetzing,* its epiphytes, and epiphyte grazers. Oecologia 85: 572-580.

Dodds, W.K. 2003. Misuse of inorganic N and soluble reactive P concentrations to indicate nutrient status of surface waters. J. N. Am. Benthol. Soc., 2003, 22(2):171–181.

Dodds, W.K., V.H. Smith, and B. Zander. 1997. Developing nutrient targets to control benthic chlorophyll levels in Streams: A case study of the Clark Fork River. Water Res. 31(7): 1738-1750.

Dodds, W.K., V.H. Smith, and K. Lohman, 2006. Erratum: Nitrogen and phosphorus relationships to benthic algal biomass in temperate streams. Can. J. Fish. Aquat. Sci. 63: 1190-1191.

Droop, M.R. 1974. The nutrient status of algal cells in continuous culture. J.Mar.Biol.Assoc. UK 54:825-855.

EPA (Environmental Protection Agency). 2001. Nutrient criteria technical guidance manual estuarine and coastal marine waters. Office of Water. EPA-822-B-01-003.

Eppley, R. W. 1972. Temperature and phytoplankton growth in the sea. Fishery Bulletin. 70(4): 1063-1085.

Fritsch, F.E., 1906. Problems in aquatic biology with special reference to the study of algal periodicity. New Phytol. 5: 149-169.

Goldman, J.C., J. J. McCarthy, and D.G. Peavey, 1979. Growth rate influence on the chemical composition of phytoplankton in oceanic waters. Nature 279: 210-215.

Horner, R.R., Welch, E.B., and R.B. Veenstra. 1983. Development of nuisance periphytic algae in laboratory streams in relation to enrichment and velocity. In Periphyton of freshwater ecosystems, Ch. 16, (The Hague: Dr. W. Junk Publishers): 121-134.

Hynes, H.B.N., 1966. The biology of polluted waters. Liverpool University Press,

Hynes, H.B.N. 1969. The enrichment of streams. In Eutrophication: Causes, consequences, correctives, Rohlich, G. A., Ch. III, (Washington, D.C.: National Academy of Sciences): 188-196.

Kannel, P.R., Lee, S., Kanel, S.R., Lee, Y.S., and K.H. Ahn. 2006. Application of QUAL2Kw for water quality modeling and dissolved oxygen control in the river Bagmati. Environ. Monit. Assess. 125: 201-217.

Klausmeier, C.A., Litchman, E., Daufresne, T. and S. A. Levin. 2004. Optimal nitrogen-to-phosphorus stiochiometry of phytoplankton. Nature. 429(2004): 171-174.

Lambing, J.H., and T.E. Cleasby, 2006. Water quality characteristics of Montana streams in a statewide monitoring network, 1999-2003. U.S. Geological Survey Scientific Investigations Report 2006-5046, 149 p.

Lohman, K., Jones, J.R. and B.D. Perkins. 1992. Effects of nutrient enrichment and flood frequency on periphyton biomass in northern Ozark streams. Can. J. Fish. Aquat. Sci. 49: 1198-1205.

Meeus, J. 1999. Astronomical algorithms. Second edition. Willmann-Bell, Inc. Richmond, VA.

McIntire, C.D. 1973. Algal dynamics in laboratory streams: A simulation model and its implications. Ecol. Monogr., 43, 399–420.

Mills, W.B., Bowie, G.L.,Grieb, T.M., Johnson, K.M., and R.C. Whittemore. 1986. Stream sampling for waste load allocation application. Washington, D.C.: U.S. EPA Office of Research and Development. Report EPA/625/6-86/013.

Morel, F. M., and J.G. Hering. Principles and applications of aquatic chemistry. John Wiley and Sons, Inc., New York.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D. and T.L.Veith. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE. 50(3): 885–900

NRCS. 2009. Phase II sedimentation assessment for the Upper Missouri River Basin. USDA Natural Resources Conservation Service Nebraska, South Dakota, North Dakota, Montana, and Wyoming. In Cooperation with Missouri Sedimentation Action Coalition

Park, S.S. and Y.S. Lee. 2002. A water quality modeling study of the Nakdong River, Korea. Ecol. Model. 152(1): 65-75.

Paschal, J. E. and D. K. Mueller. 1991. Simulation of water quality and the effects of waste-water effluent on the South Platte River from Chatfield Reservoir through Denver, Colorado. Denver, Colorado. Water Resources Investigation Report. Report WRI-91-4016.

Peterson, David A. and Stephen D. Porter. 2002. Biological and chemical indicators of eutrophication in the Yellowstone River and major tributaries during August 2000. Washington, DC: National Water Quality Monitoring Council.

Rhee, G.Y. 1973. A continuous culture study of phosphate uptake, Growth rate and polyphosphate in *Scenedesmus Sp.* J. Phycol. 9: 495-506.

0015259

Rhee, G.Y. 1978. Effects of N:P atomic ratios and nitrate limitation on algal growth, cell composition, and nitrate uptake. Limnol. Oceanogr. 23(1): 10-25.

Rier, S.T. and R.J. Stevenson. 2006. Response of periphytic algae to gradients in nitrogen and phosphorus in streamside mesocosms. Hydrobiologia 561:131–147.

Rutherford, J.C., Scarsbrook, M.R., and N. Broekhuizen. 2000. Grazer control of stream algae: modeling temperature and flood effects. J. Environ. Eng.-ASCE. 126(4):331-339.

Shuter, B. J. 1978. Size dependence of phosphorus and nitrogen subsistence quotas in unicellular microorganisms. Limnol. Oceanogr. 26(6): 1248-1255.

Smith, E.L. 1936. Photosynthesis in relation to light and carbon dioxide. Proc. Natl. Acad. Sci. 22:504-511.

Steele, J.H. 1962. Environmental control of photosynthesis in the sea. Limnol. Oceanogr. 7:137-150.

Streeter, H.W. and E.B. Phelps. 1925. A study of the pollution and natural purification of the Ohio River. III. Factors concerned in the phenomena of oxidation and reaeration. U.S. Health Service. Report Bulletin No. 146.

Stevenson, R. J., M. L. Bothwell, and R. Lowe. 1996. Algal ecology - Freshwater benthic ecosystems, San Diego, CA: Acaemic Press.

Stevenson, R.J., Rier, S.T., Riseng, C.M., Schultz, R.E., and M.J. Wiley. 2006. Comparing effects of nutrients on algal biomass in streams in two regions with different disturbance regimes and with applications for developing nutrient criteria. Hydrobiologia. 561:149-165.

Suplee, M.W. and V. Watson. 2012. Scientific and technical basis of the numeric nutrient criteria for Montana's wadeable streams and rivers – Addendum 1. Helena, MT: Montana Department of Environmental Quality.

Suplee, M. W., Watson, V., Teply, M.E. and H. McKee. 2009. How green is too green? Public opinion of what constitutes undesirable algae levels in streams. J. Am. Water Resour. As. 45(1): 123-140.

Suplee, M.W., Watson, V., Dodds, W.K., and C. Shirley. 2012. Response of algal biomass to large-scale nutrient controls in the Clark Fork River, Montana, United States. J. Am. Water Resour. As. 48(6): 1752-1688.

Thomann, R.V. 1998. The future golden age of predictive models for surface water quality and ecosystem management. J. Envir. Engin.-ASCE. 124(2): 94-103.

Turner, D.F., Pelletier, G.J. and B. Kasper. 2009. Dissolved oxygen and pH modeling of a periphyton dominated, nutrient enriched river. J. Environ. Eng.-ASCE. 135(8): 645-652.

USGS (United States Geological Survey), 2004. Water-quality assessment of the Yellowstone River Basin, Montana and Wyoming—Water quality of fixed Sites, 1999-2001. Scientific Investigation Report 2004-5113, 82 p.

Uehlinger, U. 1991. Spatial and temporal variation of the periphyton biomass in a prealpine river (Necker, Switzerland). Arch. Hydrobiol. 123(2): 219-237.

Vance, L., Stagliano, D., and G.M. Kudray. 2006. Watershed assessment of the middle Powder Subbasin, Montana. Prepared for: Montana State Office Bureau of Land Management, Billings, MT by Montana Natural Heritage Program, Natural Resource Information System, Montana State Library.

Welch., E.B., Quinn, J.M., and C.W. Hickey. 1992. Periphyton biomass related to point-source nutrient enrichment in seven New Zealand streams. Wat. Res. 26(5): 669-675.

Whiting, P. J., Matisoff, G., Fornes, W., and F. M. Soster. 2005. Suspended sediment sources and transport distances in the Yellowstone River Basin. Geol Soc. Am. Bull. 117(3-4): 515-529.

Whitton, B.A., 1970. Review Paper: Biology of *Cladophora* in freshwaters. Wat. Res. 4: 457-476.